AMENDMENTS TO THE CLAIMS:

Claims 1-52 (Cancelled)

53.   (Currently Amended) A method for a data processing system to efficiently identify at least one dataset from a collection of datasets according to a query containing information indicative of desired datasets, wherein each dataset is a document and includes one or more data points and each data point corresponds to at least one of a word, a phrase, and a sentence, ~~a color, a typography, a punctuation, a picture, and a character string~~, the method comprising the machine-executed steps:

for each dataset, constructing a semantic vector ~~for~~ representing each dataset;

receiving the query containing information indicative of desired datasets;

for the query, constructing a semantic vector ~~for~~ representing the query;

~~comparing the semantic vector for the query to the semantic vector of each dataset;~~

selecting datasets ~~whose semantic vectors are closest in distance to~~based on a distance between the semantic vector for the query and the semantic vector of each dataset; and

~~outputting~~displaying information of the selected datasets to be corresponding to the desired datasets identified in the query;

wherein:

the query or each of the datasets includes at least one data point; and

the semantic vector for the query or each of the datasets is constructed by the steps of:

for each data point, identifying a relationship between each data point and multiple predetermined categories corresponding to dimensions in the semantic space;

determining the significance of each data point with respect to the multiple predetermined categories according to a predetermined formula, ~~wherein the significance represents a relative~~

~~strength of each data point relative to each of the predetermined particular categories, or a degree of relevance of each data point relative each of the predetermined particular categories;~~

for each data point, constructing a semantic vector ~~for~~ representing each data point, wherein each semantic vector has dimensions equal to the number of multiple predetermined categories and represents the ~~relative strength~~significance of its corresponding data point with respect to each of the multiple predetermined categories; and

based on the semantic vector for each of the at least one data point, form the semantic vector ~~of~~representing the query or each of the ~~datasets.~~datasets; and

wherein the significance of each data point is determined by calculating a probability distribution of each data point occurring in each predetermined category and a probability distribution of the data point's occurrence across all predetermined categories.

54.   (Original) The method of Claim 53, wherein the datasets correspond to documents and the query is a natural language query.

55.   (Cancelled)

56.   (Original) The method of Claim 53, further comprising a step of clustering the selected datasets in real time.

57.   (Currently Amended) A method for efficiently identifying data points in a semantic lexicon related to a dataset, wherein the dataset is a document and includes one or more data points and each data point corresponds to at least one of a word, a phrase, and a sentence, ~~a typography, a punctuation, and a character string,~~ the method comprising the machine-executed steps:

constructing a semantic vector ~~for~~representing the dataset;

~~comparing the semantic vector for the dataset to a semantic vector of each of the data points~~

~~in the semantic lexicon;~~

selecting data points ~~whose semantic vectors are closest in distance to~~based on a distance

between the semantic vector for the dataset and the semantic vector of each data point; [[and]]

~~associating~~identifying said selected data points to ~~said dataset;~~be related to the dataset; and

displaying a result of the identifying step.

wherein:

the semantic vector for the dataset is constructed by the steps of:

for each data point, identifying a relationship between each data point and multiple

predetermined categories corresponding to dimensions in the semantic space;

determining the significance of each data point with respect to the multiple predetermined

categories according to a predetermined formula;~~, wherein the significance represents a relative~~

~~strength of each data point relative to each of the predetermined particular categories, or a degree of~~

~~relevance of each data point relative each of the predetermined particular categories~~

constructing a semantic vector ~~for~~representing each data point, wherein each semantic

vector has dimensions equal to the number of multiple predetermined categories and represents

the ~~relative strength~~significance of its corresponding data point with respect to each of the

multiple predetermined categories; and

based on the semantic vector ~~for~~representing each of the at least one data point, form the

semantic vector of the ~~dataset.~~dataset; and

wherein the significance of each data point is determined by calculating a probability

distribution of each data point occurring in each predetermined category and a probability

distribution of the data point's occurrence across all predetermined categories.

58. (Original) The method of Claim 57, wherein the dataset is a document and the data points are words.

59. (Original) The method of Claim 57, wherein the dataset is a natural language query in a search system and the data points are words.

Claims 60-64 (Cancelled)

65. (Currently Amended) A system for identifying at least one data set from a collection of datasets according to a query containing information indicative of desired datasets, wherein each dataset is a document and includes one or more data points and each data point corresponds to at least one of a word, a phrase, and a sentence, a color, a typography, a punctuation, a picture, and a character string, the system comprising:

a computer configured to:

construct a semantic vector for representing each dataset;

receive the query containing information indicative of desired datasets;

construct a semantic vector for representing the query;

compare the semantic vector for the query to the semantic vector of each dataset;

select datasets whose semantic vectors are closest in distance to based on a distance between the semantic vector for the query and the semantic vector of each dataset; and

output display information of the selected datasets to be corresponding to the desired datasets identified in the query;

wherein:

the query or each of the datasets includes at least one data point; and

the semantic vector for the query or each of the datasets is constructed by the machine-executed steps of:

for each data point, identifying a relationship between each data point and multiple predetermined categories corresponding to dimensions in the semantic space;

determining the significance of each data point with respect to the multiple predetermined categories according to a predetermined formula;, wherein the significance represents a relative strength of each data point relative to each of the predetermined particular categories, or a degree of relevance of each data point relative each of the predetermined particular categories

constructing a semantic vector for representing each data point, wherein each semantic vector has dimensions equal to the number of multiple predetermined categories and represents the relative strengthsignificance of its corresponding data point with respect to each of the multiple predetermined categories; and

based on the semantic vector for each of the at least one data point, form the semantic vector of the query or each of the datasets.datasets; and

wherein the significance of each data point is determined by calculating a probability distribution of each data point occurring in each predetermined category and a probability distribution of the data point's occurrence across all predetermined categories.

Claims 66-70 (Cancelled)

71. (Currently Amended) A tangible computer-readable medium carrying one or more sequences of instructions for efficiently identifying at least one data set from a collection of datasets according to an query containing information indicative of desired datasets, each dataset being a document and including one or more data points and each data point corresponding to at least one of

a word, a phrase, and a sentence, ~~a color, a typography, a punctuation, a picture, and a character string,~~ wherein execution of the one or more sequences of instructions by one or more processors causes the one or more processors to perform the steps of:

constructing a semantic vector ~~for~~ representing each dataset;

receiving the query containing information indicative of desired datasets;

constructing a semantic vector for the query;

~~comparing the semantic vector for the query to the semantic vector of each dataset;~~

selecting datasets ~~whose semantic vectors are closest in distance to~~ based on a distance between the semantic vector for the query and the semantic vector of each dataset; and

~~outputting~~ displaying information of the selected datasets to be corresponding to the desired datasets identified in the query;

wherein:

the query or each of the datasets includes at least one data point; and

the semantic vector for the query or each of the datasets is constructed by the steps of:

for each data point, identifying a relationship between each data point ~~and~~ and multiple predetermined categories corresponding to dimensions in the semantic space;

determining the significance of each data point with respect to the multiple predetermined categories according to a predetermined formula;~~, wherein the significance represents a relative strength of each data point relative to each of the predetermined particular categories, or a degree of relevance of each data point relative each of the predetermined particular categories~~;

constructing a semantic vector ~~for~~ representing each data point, wherein each semantic vector has dimensions equal to the number of multiple predetermined categories and represents

the ~~relative strength~~significance of its corresponding data point with respect to each of the

multiple predetermined categories; and

based on the semantic vector for each of the at least one data point, form the semantic vector

of the query or each of the ~~datasets.~~datasets; and

wherein the significance of each data point is determined by calculating a probability

distribution of each data point occurring in each predetermined category and a probability

distribution of the data point's occurrence across all predetermined categories.

Claims 72-75 (Cancelled)